全国第七届研究生数学建模竞赛

题 目　　神经元的形态分类和识别

摘　　要：

本文的工作主要包括以下方面：

1.提取神经元样本的空间几何属性，先进行属性选择，得出干的数目、最小直径、比较扩大角度平均值和段路径长度最大值为 4 个最主要的特征属性。然后用朴素贝叶斯模型依据选择出来的属性对神经元进行分类，最后用 10 折交叉验证方法对结果验证，取得了 96%的正确率。

2.在已有附录数据的基础上继续搜集数据作为训练样本，利用演化算法中的遗传规划模型，通过进化、选择等操作对附录 B 的数据进行预测。结果中有运动神经元 7 个、普肯野神经元 2 个、锥体神经元 4 个（包含 1 个变异）、中间神经元 2 个、感觉神经元 3 个和一个新的分类（毛线状神经元）。

3.改进已有的期望最大化(EM)聚类模型，加入双阈值终止条件和类间可分性判定依据，并将 neuromorpho.org 网站所有的 5436 组神经元数据分为 7 类，按照新类别的几何特征，尝试将其命名为多干放射状神经元、普通神经元、多干长须神经元、毛线状神经元、多干中须神经元、珊瑚状神经元和多极中间神经元。

4.根据已有的贝叶斯分类模型，区分不同物种的同类神经元。不仅对题目中猪和鼠的普肯野神经元进行了区分，还对搜集的猫和鼠的脊髓运动神经元、猴子和人类的椎体神经元进行了分组对比试验，分组试验结果精度不低于为 97.5%。

5.用分类模型取代常规的回归模型进行神经元的成长预测。根据同类神经元的大量样本聚类，把神经元的生长过程分为 6 个阶段。分析神经元的生长过程的几何特征的变化，再通过分类模型，可以确定 80%以上的神经元所处生长阶段，从而达到预测下一个或几个生长阶段的目的。

[关键词]： 属性选择　贝叶斯分类　EM 聚类　10 折交叉验证

参赛队号 10491004

队员姓名 余超、曾文聪、韩增新

参赛密码

（由组委会填写）

中山大学承办

目 录

神经元的形态分类和识别

1 问题重述

大脑是生物体内结构和功能最复杂的组织，其中包含上千亿个神经细胞（神经元）。人类脑计划（Human Brain Project, HBP）的目的是要对全世界的神经信息学数据库建立共同的标准，多学科整合分析大量数据，加速人类对脑的认识。

如何识别区分不同类别的神经元，这个问题目前科学上仍没有解决。生物解剖区别神经元主要通过几何形态和电位发放两个因素。其中，神经元的空间几何形态的研究是人类脑计划中一个重要项目。目前，该项目已经提出的主要有如下5类神经元，分别是运动神经元、普肯野神经元、锥体神经元、极中间神经元（双极中间神经元、三极中间神经元、多极中间神经元）、感觉神经元。它们从空间形态上互相差别比较显著。

本问题只考虑神经元的几何形态，研究如何利用神经元的空间几何特征，通过数学建模给出神经元的一个空间形态分类方法，将神经元根据几何形态比较准确地分类识别。

要解决的问题可以描述为如下5个：

1)根据上述提供的神经元空间几何数据建立神经元分类模型。即给出一种神经元空间形态的分类方法，根据神经元的空间几何特征将该神经元划分到某类别中。

2)根据1)中得出的分类模型，识别附录B中的神经元类别，根据识别信息，检测是否含有除上述5类外的其他神经元类型，以决定是否有必要引入或定义新的神经元名称。

3)和1)不同的是，这一问可以通过聚类的方法，通过大量神经元的空间几何信息，建立完整的神经元空间形态分类模型。分析聚类结果，总结各类神经元的形态特征，为生物学家提供命名建议。

4)用模型区分不同物种相同类型的神经元在空间几何特征上有何异同。

5)尝试预测神经元形态的生长变化，并分析生长变化对分类模型的影响。

2 问题分析与解题思路

本题是一个结合属性选择、分类、聚类、预测等多个统计学习方面的综合问题。该题的关键点有如下几个：1.大规模数据集的预处理。通过软件或编程计算，得出一定数量的去量纲化的特征属性。2.建立基于特征属性的分类模型。通过模型的特性，可以分析总结出各类神经元的空间几何特征，并据此分类。3.对于超出已有类别的特殊数据，归纳出其特征属性的取值，总结其特征。4.对于所有数据，在类别未知的情况下，归纳出特征明显的多类，并总结这些特征。5.提取同一类别神经元在不同物种内的特征。6.统计回归神经元生长的模型，并依次进行预测。7.确保分类模型对生长变化的神经元外形特征的包容性。

3 部分符号说明

A_i：第 i 组属性名，i = 0,1,…,N。每个属性的名称详见附录 1。

C_i：第 i 类神经元，i = 0,1,…,6。第 0 至 6 类依次代表运动神经元、普肯野神经元、椎体神经元、双极中间神经元、三极中间神经元、多级中间神经元和感觉神经元。

T_i：第 i 组训练样本，i = 1，2，…，N。N 表示样本的个数。

X_i：第 i 组数据的属性，i = 1，2，…。

CL_i：类标记，i = 1，2，……，N。N 表示样本的个数。

ZA_i：标准化后的属性，i = 1，2，……，N。N 表示样本的个数。

4 数据搜集及预处理

4.1 数据搜集

这里用到的所有数据全部来自国外网站，网址在题目中已经给出。第一问的数据为附录 A、C 的数据，不需要搜集。第二问需要自己找数据。为了使结果更加精确，每个类别的数据都找了 10 组，除其中一组数据只有 6 个，因此数据总量为 66 组。第三问题目也没有直接给出数据。为了区分神经元的类别，用到了题目提供的第一个网站中的 5436 组数据。第四问用到了题目给出的 6 组数据，同时也搜集了其 40 组数据进行对比。第五问用到的是人类椎体神经元数据，搜集到 1908 组数据。

4.2 数据预处理

题目通过 swc 文件给出了神经元的三维结构形态，我们可以根据题目给出的文件刻画完整的神经元空间几何特征。由于题目中的网站提供的其他软件或是无法运行(如 L-Measure)，或是运行结果和题目给出的数据稍有差异(Neuromantic)，所以我们决定通过软件 Neuron 提取神经元的一系列属性。

具体方法如下：

进入文件根目录，输入命令“MorphometricsList.pl 文件相对路径”。

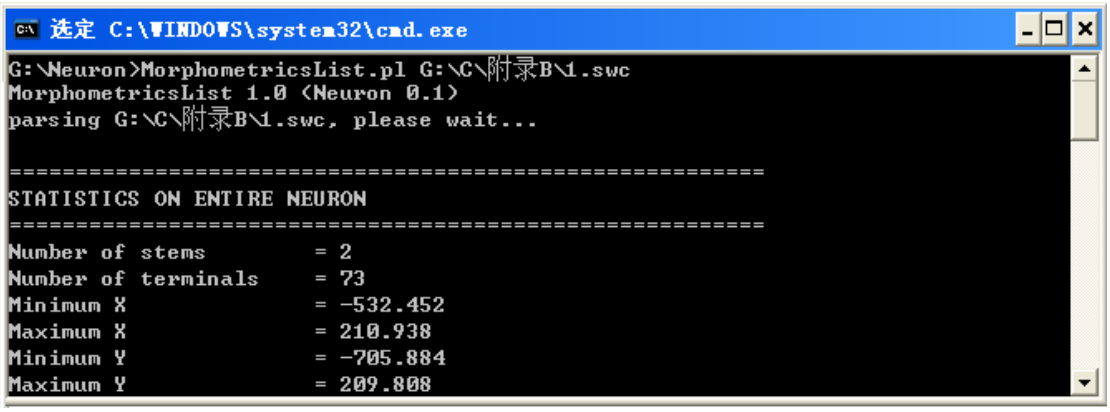


图 4-1 软件 Neuron 运行的部分截图

提取命令行生成的数据，使用批处理命令将数据导入到 excel 文件中。

可以看出软件生成了一些属性，如 Number of stems, Number of terminals 等。经过统计分析，处理后的结果包含 78 个特征属性，记为 A0 ,A1，…，A77。其中有 4 个特征属性(分别是 Minimum branch order，Min path distance，Min eucl. Distance，Min comp. length)对于全体数据集均为 0，可认为是无用属性直接删除。剩下 74 个属性可以完全刻画神经元的几何形态。

经过筛选后的部分数据如下图所示：

	A	B	C	D	E	F	G	H
1	A0	A1	A2	A3	A4	A5	A6	A7
2	2	73	-532.452	210.938	-705.884	209.808	-180	900.64
3	12	187	-1206	1306	-1217	1108	-1479.38	1050
4	13	140	-1564.31	1222.316	-1424.02	655.7981	-1085.71	1579.727
5	11	127	-1520.45	787.466	-1766.38	1518.845	-2689.75	1554.686
6	2	19	-153.18	-1	-226.23	-1	26.71	81.19
7	2	7	-129.15	-1	-266.58	-1	-48.05	31.76
8	2	58	-482.25	293.44	-181.928	329.989	-2.832	23.6
9	2	35	-329.414	222.454	-210.272	336.698	-26.597	26.456
10	2	89	3.596	252.914	5.428	114.762	5.882	20.587
11	2	62	12.44	133.33	2.9	190.45	4.412	29.41
12	8	121	-335.72	533.56	-300	419.11	-463.68	320.15
13	2	114	-727.151	250.665	-359.273	191.813	-226	268.64
14	7	54	-64.73	425.74	-234.61	217.52	-133.1	139.15
15	6	2293	-1087.44	1952.8	-401.745	2480.55	-301.155	111.318
16	1	418	-83.5	201	-217.5	28.5	-9	15.5
17	1	473	-125.5	134.5	-306	0	-2	34.5

图 4-2 预处理后的数据

过多的属性包含大量的噪声属性，会对分类结果的正确性产生影响，属性的筛选直接影响了分类的准确性。因此，在进行分类前会对这些属性进行特征提取，选择最具代表性的几个属性进行分类。

5 问题 1——属性选择

5.1 问题分析

该问题是一个需要基于特征选择的分类模型问题。由于预处理后的数据含有

70 多个神经元空间几何信息，直接以这些几何信息作为属性进行分类的效果差。太多的属性会产生很多噪声，相互干扰，影响分类的结果。这样，就需要在分类前进行特征选择，筛选出具有代表性的几个属性用来刻画附录 C 中 5 类神经元的几何特征。

为了区分不同类别的神经元，还需要通过选择出来的属性，建立模型对附录 C 的数据进行分类。我们可以根据自己的模型，给出一个神经元空间形态的分类方法。

5.2 模型建立

5.2.1 特征选择

在特征（属性）评价选择体系中，现在已有几十种各式各样复杂的方法，如主成分分析、因子分析、TOPSIS 分析、突变理论、ELECTRE 等等[1]，这些评价方法大致可以分为线性评价方法和非线性评价方法，前者采取一定的方法给指标体系赋权，然后进行加权汇总，后者相对复杂，原理不同方法不同，其特点是指标体系和评价结果之间的关系呈现非线性。

特征的选择问题，归根结底可以理解为在所有属性集上的组合问题。从组合数学的角度来看，78 个属性的特征提取（降维）问题，是在一个 2^{78} 的解空间中搜索最优子集的问题。

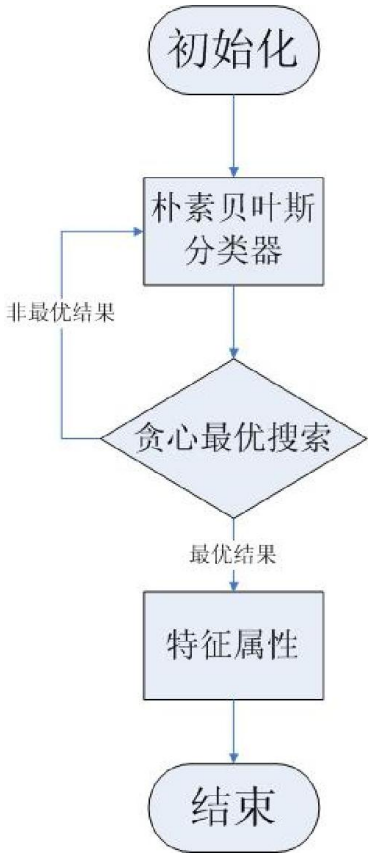


图 5-1 属性选择过程

问题 1 的数据量大，原始特征数也非常多。传统的采取统计的方法计算出评价的指标，面对大量的特征，使用统计的方法计算起来往往捉襟见肘，准确率较

低且不好验证。因此，决定采用搜索加评价的模型体系，用搜索算法在解空间中搜索属性子集，用分类模型的分类准确率作为评价的标准。

盲目搜索的效率很低,耗费很多时间和空间,如果我们在搜索时,能首先选择最有希望的节点，我们称这种搜索为“启发式搜索”或“信息搜索”，如最优搜索,最优搜索也就是判断哪个为最有希望的节点有序搜索。最优搜索的方法很多，介于本题数据较为简单，本文中我们采用的是贪心最优搜索算法。

属性选择的整体过程如图 5-1 所示。

算法 5-1 最好优先选择算法

算法描述：

Step1： OPEN = 初始集合。

Step2： 当 OPEN 不为空时执行 Step3-Step7。

Step3： 选择最好的路径点加入到 OPEN 中。

Step4： 建立下一步节点集合。

Step5： 对于每一个下一步节点执行 Step6-Step7。

Step6： 如果这个节点没有初始化：用朴素贝叶斯模型评价它，将它加入到 OPEN 中，并记录其上一步节点。

Step7： 否则，如果新路径比上一个路径好，就更换掉它的上一步节点。

Step8： 输出所选的节点（即选择的属性）。

经过计算，我们淘汰了 74 个属性中的 70 个。由此看来，提取 4 个属性是最好的。

剩下 4 个最具有代表性的特征数据 A_0、A_{15}、A_{37} 和 A_{67}，如下表所示：

表 5-1 选择的属性及其物理含义

属性编号	A_0	A_{15}	A_{37}	A_{67}
物理意义	干的数目	最小直径	“比较、扩大、角度”平均值	“段、路径、长度”最大值

根据选择的属性，可以得到下面的分布图。

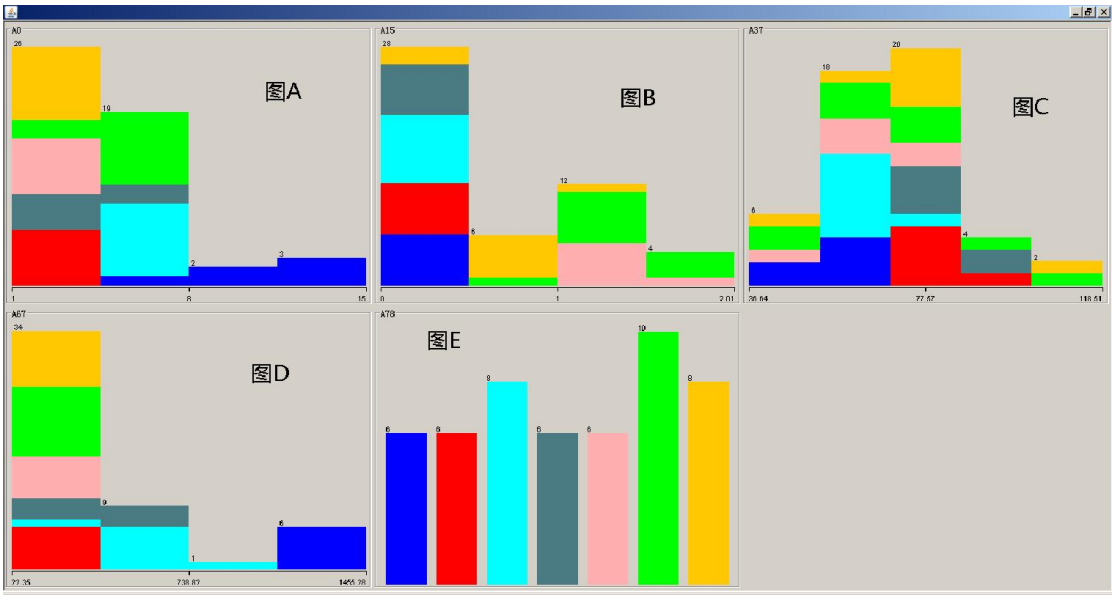


图 5-2 属性分布图

属性分布图可以形象的表现出各类神经元不同属性的取值范围。

首先，图 E 中标识了各个类别所代表的颜色。蓝色代表 C_0，红色代表 C_1，青

色代表 C_2,灰色代表 C_3，粉色代表 C_4，绿色代表 C_5,黄色代表 C_6。不同颜色的色块表示相应的类别。

图A到D代表了不同属性上的不同类别数值的分布情况,横坐标代表其数值。其中图 A 表示 A_0属性，图 B 表示 A_{15}属性，图 C 表示 A_{37}属性，图 D 表示 A_{67}属性。

如果在一个柱状图上有 n 个颜色的色块叠加，那么表示，如果某一神经元对应属性的取值在这段区间中，那么这个神经元可能是 n 种色块代表的类别中的一种。当 n=1 时，这个属性可以作为判断是否为此类别的标准。

例如图 A 所示，当 A_0属性大于 8 时，我们可以直接判定这个个体为 C_0类。

但一般情况下 n 不等于 1，所以，需要多个属性，通过多个取值区间的判断，来推算出类别。

例如某数据的 A_{37}属性在 53.052 到 69.464 之间（图 C 的第二个柱子上，可能是 $C_0, C_1, C_2, C_4, C_5, C_6$），且其 A_{67}属性在 380.585 到 738.82 之间（图 D 的第二个柱子上,可能是 C_2, C_3），那么我们取其交集，可以确定，这个数据是 C_2类。

5.2.2 构造朴素贝叶斯分类器

构造分类器需要有一个训练实例集作为输入，训练实例集中的每一个实例可描述为属性和类组成的向量，具体形式为$\langle a_1, a_2, ..., a_m, c\rangle$，其中，$a_1, a_2, ..., a_m$分别为 m 个属性的取值，c 为该训练实例的类标记。

分类器的构造方法很多，常见的有贝叶斯方法、决策树方法、基于粗糙集的方法、基于模糊集的方法等等。其中，贝叶斯方法以其独特的不确定性知识表达形式、丰富的概率表达能力、综合先验知识的增量学习特性等成为众多方法中最为引人注目的焦点之一。因此，我们采用朴素贝叶斯方法进行分类。

根据朴素贝叶斯分类器的分类公式：

$$c(x) = \arg\max P(c) \prod_{\substack{i=1 \\ c \in C}}^{m} P(a_j \mid c) \tag{5-1}$$

$$P(c) = \frac{\sum_{i=1}^{n} \delta(c_i, c)}{n} \tag{5-2}$$

其中，$P(c)$为类别 c 在训练集中出现的概率，上述公式中，n 为训练实例个数、c_i为第 i 个训练实例的类标记，a_{ij}为第 i 个实例的第 j 个属性值，$\delta(c_i, c)$为一个二值函数，当c_i=c 时取 1，其他情况时为 0。

本题中所有的属性都是数值属性，在处理数量值时，通常把它们假设成拥有“正态”的概率分布形式。

$$P(a_j \mid c) = \frac{1}{\sqrt{2\pi\sigma_j}} e^{-\frac{(a_j - \mu_j)^2}{2\delta_j^2}} \tag{5-3}$$

其中，

$$\mu_j = \left(\sum_{i=1}^{n} a_{ij}\right) \Big/ n \tag{5-4}$$

$$\sigma_j = \sqrt{\left(\sum_{i=1}^{n} (a_{ij} - \mu_j)^2\right) \Big/ (n-1)} \tag{5-5}$$

朴素贝叶斯网络的模型的建立就是贝叶斯网络训练的过程。通过训练实例集计算先验概率，即用公式（5-4）（5-5）计算出所有属性的先验概率，表 5-1 是问题 1 所有属性的先验概率。在判定新实例属于哪种类型时，给定一组实例 $\langle a_1, a_2, ..., a_m \rangle$，根据式（5-1）计算出各个已知类别的 $P(C_i)$ 的概率值，拥有最大概率的类型即可被认为是该实例的类型。

表 5-1 C_0-C_6的先验数据

名称	比例	属性	平均值	标准差
C_0	0.12	A0	11.4722	2.174
		A15	0.2047	0.0416
		A37	53.1639	5.1567
		A67	1273.711	132.0132
C_1	0.12	A0	1.3611	0.4348
		A15	0.0372	0.0526
		A37	80.7409	4.5842
		A67	34.828	11.1254
C_2	0.16	A0	6.5625	0.9998
		A15	0.349	0.0369
		A37	64.6069	6.5199
		A67	529.8837	167.381
C_3	0.12	A0	3.6944	0.8017
		A15	0.335	0.0186
		A37	83.5839	5.5711
		A67	348.2804	72.4434
C_4	0.12	A0	3.5	0.1944
		A15	1.3958	0.1248
		A37	64.2516	8.6793
		A67	189.0665	63.7166
C_5	0.19	A0	5.95	1.7655
		A15	1.4405	0.2616
		A37	70.2787	18.9612
		A67	137.322	68.3357
C_6	0.16	A0	2.1875	0.3858
		A15	0.7119	0.2732
		A37	80.3855	18.0007
		A67	78.3631	33.167

5.3 模型求解

题 1 的要求是给出一个神经元空间形态分类的方法，我们采用的是朴素贝叶斯分类器进行分类，表 5-1 给出的各个属性的先验概率就是朴素贝叶斯的模型。

模型建立后，需要对该模型的准确性进行测试，因此需要训练集数据。题 1 要求只用附录 A 和附录 C 的数据，但是附录 A 和附录 C 中的数据量比较少，如果再区分训练集和测试集会使训练实例数目更少，因此采用 10 折交叉验证的方法，将数据集分成十分，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验。每次试验都会得出相应的正确率。10 次结果的正确率的平均值作为对算法精度的估计，一般还需要进行多次 10 折交叉验证（例如 10 次 10 折交叉验证），再求其均值，作为对算法准确性的估计[4]。

在总样本中随机抽取 10%的样本作为测试，剩余的 90%数据作为训练样本，运行 10 次取平均值即为该模型的分类精度。计算公式如下：

$$Accuracy = \frac{\sum_{i=1}^{10} E_i}{N} \qquad (5\text{-}6)$$

式中 Accuracy 表示验证的精度，Ei 为每次测试被正确分类的样本个数，N 为样本总数。

经过 10 折交叉验证，得到如下结果（只举 10 个例子中的一个），详细数据请参见附录 2：

表 5-2 结果举例

编号	C_0类概率	C_1类概率	C_2类概率	C_3类概率	C_4类概率	C_5类概率	C_6类概率	实际类别	预测类别	正确率
1	0	0	0	0	0	0.03	*0.973	C6	C6	100%
2	0	0	0	*1	0	0	0	C3	C3	
3	0	0	*1	0	0	0	0	C2	C2	
4	0	0	0	0	0.01	*0.994	0	C5	C5	
5	0	*0.97	0	0	0	0	0.03	C1	C1	

注:*表示概率最大的数据，并据此判定类别。

表 5-3 部分 10 折交叉验证结果

次数	正确率	次数	正确率
第一次	100%	第六次	100%
第二次	80%	第七次	80%
第三次	80%	第八次	80%
第四次	100%	第九次	100%
第五次	100%	第十次	100%

在 50 次分类预测中，只有两次错误的判断（一次将 C4 误判成 C5，一次将 C6 误判成 C5）。分类正确率的均值可以达到 96%，因此该模型分类效果较好。

6 问题 2——样本预测

6.1 问题分析

获取了附录 B 的 19 个神经元数据（不含重复的第 9 组数据），现在要通过样本的特征属性，对这些样本的类别进行预测。不仅如此，还需要检测附录 B 中是否含有新的类别。

问题一中的朴素贝叶斯分类模型不能处理该类问题，所以需要引入一种新的分类方法。我们建立了基于遗传规划的分类模型，它通过检验是否拒绝属于某类别的假设来判定样本类别，如果所有类别假设都不成立，则表明其为未知类别的异类样本。

6.2 模型建立

基于上面的理解，关键问题在于分类的方法。我们将根据分类方法建立模型，预测样本的类别。

除第 4 类神经元只搜集到 6 个样本外，我们对其他每类神经元均搜集了 10 个训练样本。设训练样本 T_i (i = 1，2，……，N)共 N 组数据，每一组数据含属性 X_j(j = 1，2，……)和类标记 CL_i(i = 1，2，……，N)，即我们知道样本属于哪一个类。

为了保证计算结果不受量纲的影响，我们采用归一化的方法对数据进行处理。

归一化公式：

$$X = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{6-1}$$

X 为处理后的结果，x_i 表示待处理的样本的第 i 个属性，x_{min} 是该属性的最小值，x_{max} 表示该属性的最大值。经过公式 4-1 的处理后的结果如表 4-2 所示。

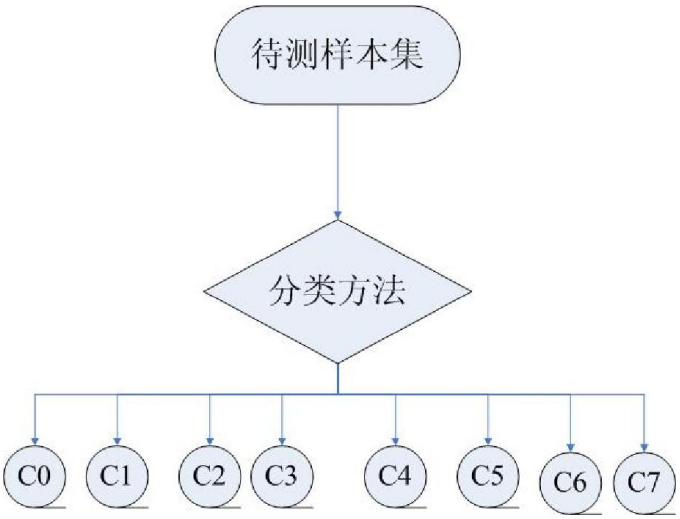


图 6-1 样本预测模型

我们假设目前存在的 7 个类别分别为 C_i（i = 0，1，……，6），并假设存在独立于 7 类之外的类别 C_7 存在。那么可以根据样本的特征，建立数学模型预测样本的类别。

如果存在独立于 7 类之外的样本，全部分到第八类；如果不存在，则第八类 C_7 中不会分到数据。至于第八类包含多少小类，则不在考虑范围之中。

根据已有的 7 类数据，建立 7 个分类器 G_m(m = 0,1，……，6)。每个分类器对每个训练样本做出判定，接受或者拒绝这个样本，根据判定结果判断分类器的好坏。每个分类器都需要对所有训练样本做出回应，这样每次训练都需要判定 7N 次。通过遗传操作更新分类器，直到产生满足下列条件的分类器。我们现在要做的是这样选择出这样的分类器。

$$C_i >= 0, Cm(m \neq i) < 0, i = (0,1,...,6) \qquad (6\text{-}2)$$

理想的情况是，对于属于第一类的样本，我们只要分类器 C_0 受它，其他的分类器表示拒绝，并以此类推。

接下来训练分类器，这里给出一个分类器的模型。

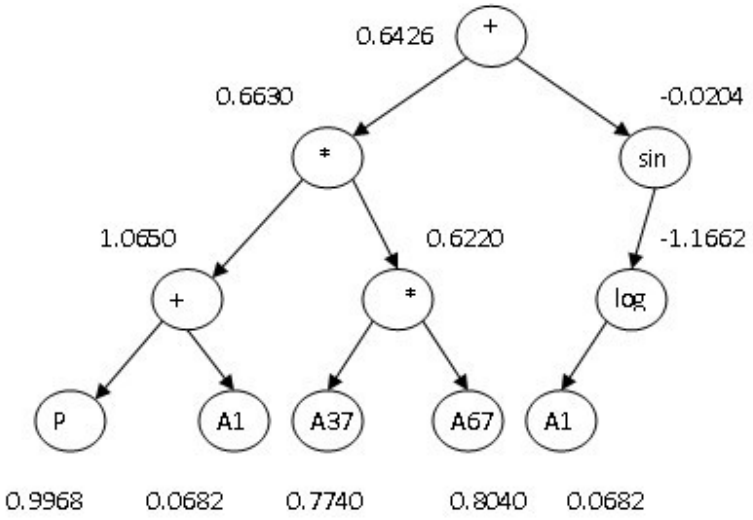


图 6-2 一个分类器的模型

分类器的选择属于遗传规划算法的范畴，包含选择、交叉、变异操作。

基本流程如下：

第 1 步：随机产生一定数量的初始群体，每个个体代表一个表达式。给定中间节点的数据集为 O={+, -, /, *, Sq, Sqrt, If, If3, >, <, !, Pow, &, |, Xor, Max, Min, Exp, Log, Sin, Cos}，而终止节点集为属性 A_i（i=1, 2, …）和任意随机常数 P。

第 2 步：计算个体的适应度，根据下面的适应度函数判别分类器的优劣程度。

$$\begin{cases} B(X_j) = 0, \text{分类器判别错误} \\ \mathrm{B}(X_j) = 1, \text{分类器判别正确} \end{cases} \qquad (6\text{-}3)$$

$$f(C_i) = \sum_{1}^{n} B(X_j)$$

并判断是否达到精度要求。若满足，输出最佳个体及其代表的最优解，并结束运算否则转向第 3 步。

这里设定精度要求：

$$f(C_i) < N * 5\% \qquad (6\text{-}4)$$

模型要求分类精度控制在 95%以内，即误判率为 5%。有时候达不到这个精度，那么需要设定一个运行时间，如果超过这个时间还没有达到要求，也必须终止运行。

第 3 步：依据适应度按照下面概率公式选择优良个体，一般用适应度高的个体覆盖适应度的低的个体，达到加速进化的目的。

$$P_i = \frac{f(C_i)}{\sum_{1}^{M} f(C_i)} \qquad (6\text{-}5)$$

P_i 表示个体被选取的概率。

第 4 步：按照一定的概率对个体进行交叉操作，生成新的个体。

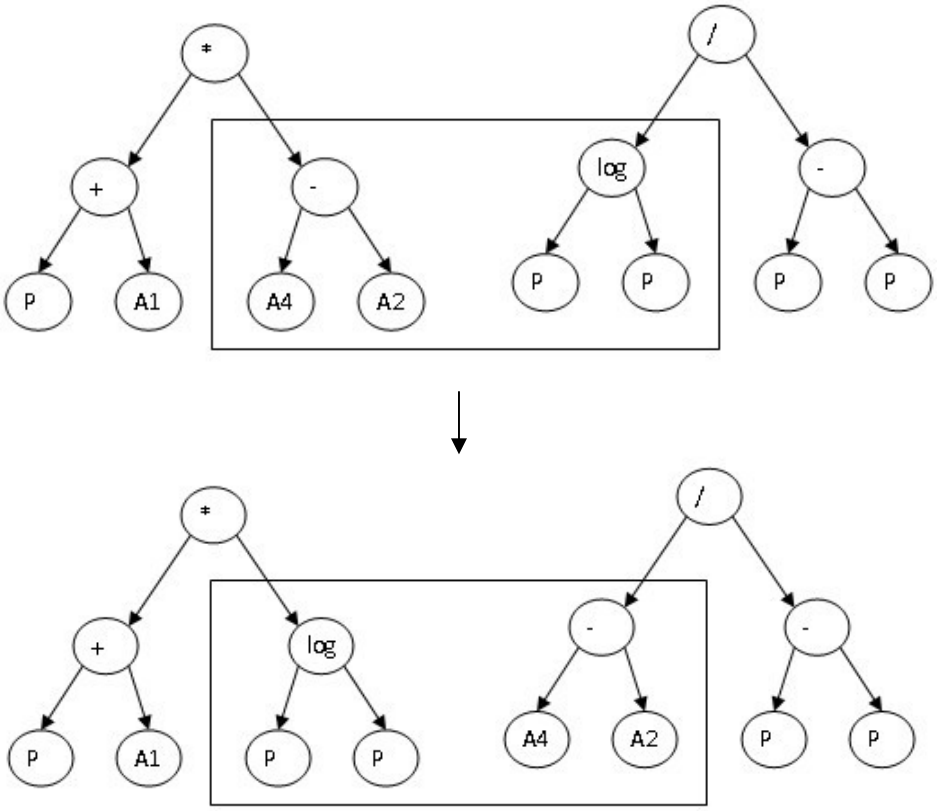


图 6-3 分类器的交叉操作

注意加框部分发生了变位，产生了新的分类器。交叉的概率可以根据问题域的不同，灵活的设定。

第 5 步：产生新一代的种群，返回到第 2 步。

一旦得到了满足条件的分类器，就可以对待测样本进行分类了。

6.3 模型求解

我们给出一组样本的处理结果，处理方法已在数据预处理一节详细说明。为

了能更好的说明问题，我们给出隶属于运动神经元的样本名为 ok_m85mod.CNG.swc 的样本。

表 6-1 样本 1 处理结果

属性	A0 （干的数目）	A1 （节点数）	……	A77 （路径距离标准差）	类标记
样本 1	7	161	……	157.7693	0

首先归一化消除量纲的影响。

表 6-2 样本 1 归一化结果

属性	A0 （干的数目）	A1 （节点数）	……	A77 （路径距离标准差）	类标记
样本 1	0.4286	0.0682	……	0.2593	0

通过遗传算法编程，尽管我们用到了多种操作符参与运算，但是我们得到的分类器结果只含有其中的两种“+”和“-”。分类器表达式最终结果如下：

C0：A13-A29 C1：A2-A60 C2：A29+A9-A16 C3：A20-A13

C4：A42-A32 C5：A41-A68 C6：A70-A1

对于样本 1 来说，A13=0.1538，A29=0.7699，利用分类器 C0 进行判定：

$$C_0=A13-A29=0.1538-0.7699=-0.6161<0$$

因此 C_0 对样本 1 的判定结果为“-1”，拒绝样本 1。但是同时可以有其他的分类器也接受样本 1，但也会出现多个分类器都接受，或者所有分类器都拒绝的情况发生。这将在后面详细讨论。据此，我们根据建立的模型来解决问题，如果一个样本被所有的分类器都拒绝，那么这个样本可以单独作为一类，即 C_8（第八类）处理。先给出计算出的分类器对样本的判定图。

表 6-3 分类器对 B 样本的判定图

B 样本	C0	C1	C2	C3	C4	C5	C6	判定结果
1	-1	-1	1	1	-1	-1	1	C2
2	-1	-1	1	-1	-1	-1	-1	C2
3	-1	1	1	-1	-1	-1	-1	C2
4	-1	-1	1	1	-1	-1	1	C2
5	-1	1	-1	-1	-1	-1	1	C1
6	-1	1	-1	-1	-1	-1	1	C1
7	1	1	-1	-1	1	-1	-1	C0
8	1	-1	-1	1	1	-1	-1	C0
10	1	-1	1	1	1	1	1	C0
11	1	-1	1	1	-1	1	1	C0
12	1	-1	1	1	-1	1	1	C0
13	-1	-1	-1	-1	-1	-1	1	C6
14	-1	-1	-1	-1	-1	-1	1	C6
15	-1	-1	-1	1	-1	-1	1	C3
16	-1	-1	1	1	-1	1	1	C5
17	-1	-1	-1	-1	-1	-1	-1	未判定
18	-1	-1	-1	-1	-1	-1	1	C6
19	1	1	1	-1	-1	1	1	C0
20	1	1	1	-1	-1	1	1	C0

其中“1”表示分类器接受样本，“-1”表示拒绝。为了正确的判定样本的类别，我们有如下定义。

定义 1：用 D_{ii} 表示真实类别为 i（1,2，……）的样本被分类器 i（1,2，……）接受的个数，S_i 表示样本 i 被其他分类器接受的个数。关联强度矩阵 SA_i 为

$$SA_i = D_{ii} /(S_i + D_{ii}) \qquad (6\text{-}6)$$

根据分类器的性能，很容易计算出相应的 D_{ii} 和 S_i 的数值，通过这些数算得的关联强度矩阵如下表所示。

表 6-4 关联强度矩阵

	SA0	SA1	SA2	SA3	SA4	SA5	SA6
强度系数	0.41	0.237	0.29	0.25	0.29	0.32	0.18

对于样本 1,有 C2、C3、C6 三个分类器表示接收。我们比较关联强度矩阵发现：SA2>SA3>SA6,根据以上判定图和关联强度矩阵，可以把样本 1 划为 C2，即椎体神经元。

把分类器和神经元的类型进行映射，得到 19 个样本的类标记如下表所示：

表 6-5 预测 B 样本的类别

类别	标记	B 组样本编号
(1)运动神经元	C0	7,8,10,11,12,19,20
(2)浦肯野神经元	C1	5,6
(3)椎体神经元	C2	1,2,3,4
(4)中间神经元	(4a)双极中间神经元 C3	15
	(4b)三极中间神经元 C4	
	(4c)多极中间神经元 C5	16
(5)感觉神经元	C6	13,14,18

我们用软件 Neuromantic 画出了所有的 B 组数据的空间三维图形后，发现了两个比较独特的神经元，如下图所示。

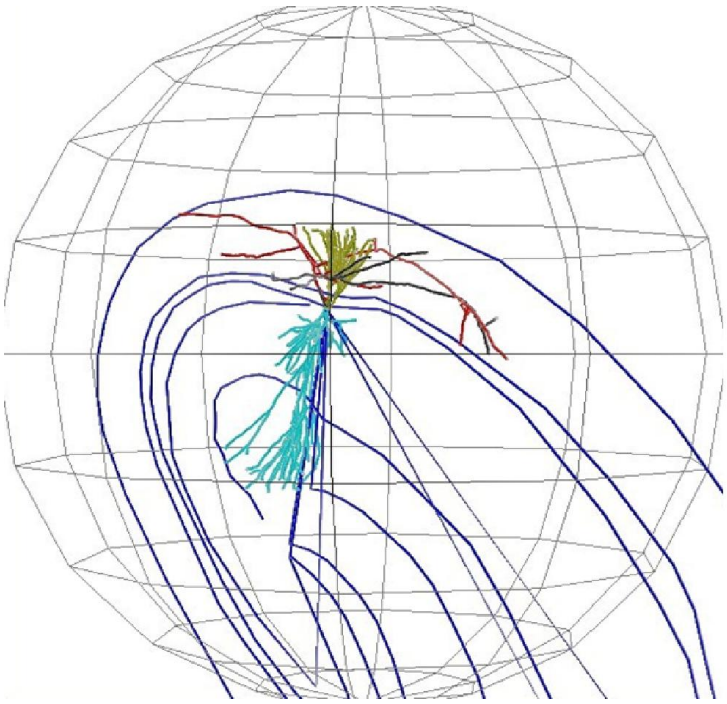


图 6-4 B 数据的第四号神经元

根据上述数学方法，我们把这组数据判定为第三类即椎体神经元。如果出现新类，我们预计会得到被所有分类器拒绝的样本，但此图并没有发生这种情况。图 X 中的神经元不仅包含椎体神经元的特征，而且还明显含有长须状结构。由于模型把它判定为椎体神经元，所以我们认为没有必要引入新的神经元定义。椎体神经元中存在一类神经元，它的特性是总面积不大，但是长宽高等数据都比较大。

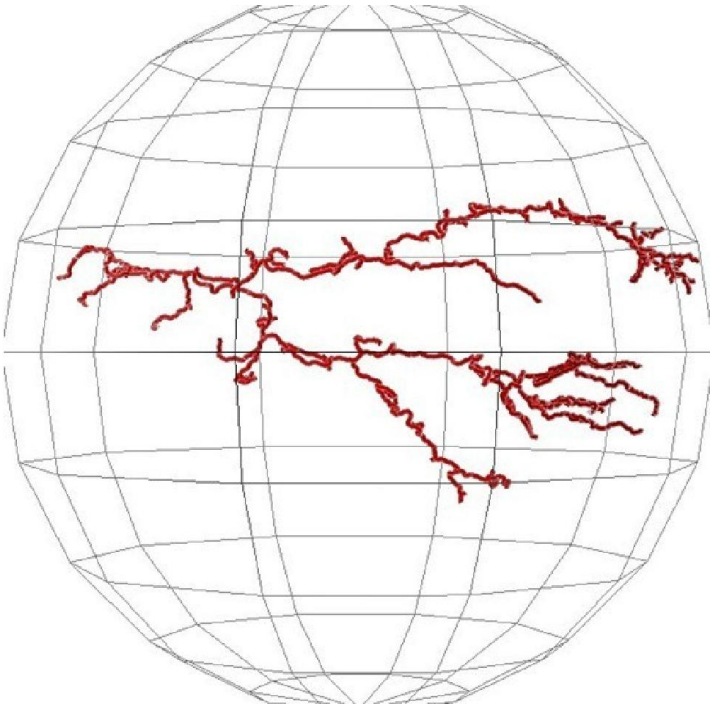


图 6-5 B 组第 17 号神经元

模型未能给出这个神经元的判定，所有类别的分类器都拒绝了这个神经元，我们成图后觉得这个神经元比较特殊，可以作为新类别的神经元。这类神经元的平均分叉扭矩大，可以通过这种形态特征单独作为一种类别的神经元。对此，可以把它判定为感觉神经元，因为感觉神经元的平均分叉扭矩比较大，也可以引入新的神经元名称。我们定义它为“毛线状神经元”。

7 问题 3——分类识别

7.1 问题分析

在未知类别的情况下对数据进行分类，只能用聚类的方法。而聚类需要大量的样本以提高模型的精度。为了保证聚类的客观性和准确性，我们先从网站下载了 5436 组数据，通过 Neuron 软件批量提取了它们的空间几何特征。然后跟它们的特征进行聚类，常用的聚类方法有统计方法、网格方法、基于密度方法、K 均值聚类、演化聚类等。

由于输入数据量大，聚类方法往往都是耗时操作。为了减少计算时间，提高模型效率，我们可以通过模型对原始数据降维，提取特征属性后再进行聚类。聚类方法给每个样本一个类别编号，标注为同一个类别编号的被划为同一类。但是我们不知道每一类具体代表什么空间几何意义。这就需要我们分析样本的特征属

性，并找出每类样本的空间几何意义，给出自己定义的命名。

7.2 模型建立

我们首先建立期望最大化（EM）模型[2]，假设某随机向量由两部分组成：可见部分 Z 和隐含部分 C，满足分布 $P(C,Z;\theta)$。训练数据由若干 Z 的实例组成 $Z=\{z_1,z_2,...,z_m\}$现在要对参数θ进行估计。为简化讨论，假设 C 是一个离散随机变量,满足分布$[\alpha_1,\alpha_2,...,\alpha_n]$，即 $p(C=i)=\alpha_i$。

θ的最大似然估计值为 $\theta^*=\arg\max P(Z|\theta)$ 或 $\theta^*=\arg\max \ln P(Z|\theta)$，但是由于 $P(Z|\theta)$ 可能非常复杂，难以进行优化。

当函数 $f(x)$ 是凸函数时，琴生不等式保证：$f(\sum\alpha_i x_i)\geq\sum\alpha_i f(x_i)$，其中：$\alpha_i\geq 0,\sum\alpha_i=1$

容易证明函数 log（x)是凸函数，所以 $\ln(\sum\alpha_i x_i)\geq\sum\alpha_i\ln(x_i)$ 于是：

$$\sum\alpha_i x_i\geq\prod\alpha_i x_i \tag{7-1}$$

将 $p=(C=i,Z|\theta)$ 简写为 $p=(C_i,Z|\theta)$ ，我们有：

$$p(Z|\theta)=\sum p(C_i,Z|\theta)=\sum\frac{p(C_i,Z|\theta)}{C_i}\alpha_i\geq\prod[\frac{p(C_i,Z|\theta)}{\alpha_i}]^{\alpha_i} \tag{7-2}$$

其中 2 的最后一个不等式应用式 1。对 2 两边取对数得到：

$$\ln p(Z|\theta)\geq\ln\prod[\frac{p(C_i,Z|\theta)}{\alpha_i}]^{\alpha_i}=\sum\alpha_i(\ln p(Z|\theta)-\ln\alpha_i) \tag{7-3}$$

记 $L(\theta)=\ln p(Z|\theta),F(\alpha,\theta)=\sum\alpha_i(\ln p(C_i,Z|\theta)-\ln\alpha_i)$。

其中：$\alpha=(\alpha_1,\alpha_2,...,\alpha_N)^T$。则公式 X-X 表明，对于$\forall\alpha,\theta$：

$$L(\theta)\geq F(\alpha,\theta) \tag{7-4}$$

可以看出 $F(\alpha,\theta)$ 是 $L(\theta)$ 的下界，期望最大化模型通过优化 $F(\alpha,\theta)$ 来实现优化 $L(\theta)$ 的目的。

假设在 t 时刻，已经有对 θ 的估计值 θ^t，根据 7-5 式，我们有：

$$L(\theta^t) \geq F(\alpha, \theta^t) \tag{7-6}$$

当$\alpha^t = p(C_i \mid Z, \theta^t)$时，公式 7-6 等号成立，即

$$L(\theta^t) = F(\alpha, \theta^t) \tag{7-7}$$

固定$\alpha = \alpha^t$，则$F(\alpha^t, \theta)$为θ的函数，对此函数进行优化。若该函数在θ^{t+1}取得最大值，则有：

$$F(\alpha^t, \theta^t) \leq F(\alpha^t, \theta^{t+1}) \tag{7-8}$$

在得到θ^{t+1}之后，固定$\theta = \theta^{t+1}$，则$F(\alpha, \theta^{t+1})$为α的函数，根据式 5，我们有：

$$L(\theta^{t+1}) \geq F(\alpha, \theta^{t+1}) \tag{7-9}$$

当$\alpha^{t+1} = p(c \mid Z, \theta^{t+1})$时，式（7-9）等号成立，即：

$$(\theta^{t+1}) = F(\alpha^{t+1}, \theta^{t+1}) \tag{7-10}$$

根据定义，α^{t+1}使得函数$F(\alpha\ , \theta^{t+1})$取最大值，所以：

$$L(\theta^{t+1}) = F(\alpha^{t+1}, \theta^{t+1}) \geq F(\alpha\ , \theta^{t+1}) \qquad \forall \alpha \tag{7-11}$$

综合(7-10)式，我们有：

$$L(\theta^{t+1}) = F(\alpha^{t+1}, \theta^{t+1}) \geq F(\alpha, \theta^{t+1}) \geq F(\alpha, \theta) = L(\theta) \tag{7-12}$$

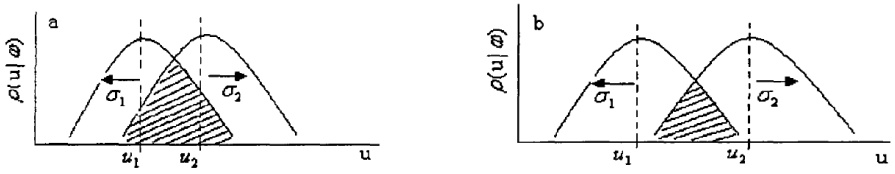
式(7-12)表明，EM 模型的每轮运行，都会使得目标函数$L(\theta)$的值增加，从而保证 EM 模型会收敛到某局部最优解。

由于 EM 模型不能保证取得全局最优解，因此为了保证结果的高效性，我们需要改变结果的终止条件。这里设置双阈值法，即迭代次数和结果精度。当模型计算结果达到精度要求后可以终止运行，或者到预先设定的次数但没有达到精度要求，这时候也应该停止。精度的公式如下：

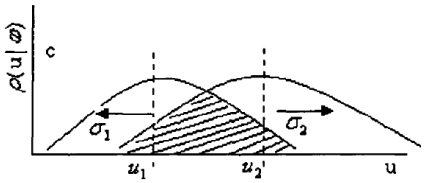
$$\sum_{j}^{M} \left| \frac{\theta_j^{(i)} - \theta_j^{(i-1)}}{\theta_j^{(i)}} \right| < T \tag{7-13}$$

这里，M 为提取的特征数，$\theta_j^{(i)}$表示第 i 次参数估计向量的第 j 个元素；T 为先设定的估计改变上限。随着迭代次数 i 的递增，式(7-13)左端量值应当递减，直至小于 T，则迭代结束。

还需要考虑一个问题，那就是用期望最大化方法聚类后，类别个数的设定。为了避免主观因素对聚类结果造成影响，我们使用“类间可分性”原理，保证分出来的类别区分度最大。



(a) 两个相互重叠的正态密度函数　　(b) 均值间的距离增大，则重叠度变小



(c) 标准差增大，则重叠度增大

图 7-1 标准差距离

上图显示了三对具有不同重叠程度的正态概率密度函数。若标准差相同（如图 7-1a、b），均值间的距离($u_1 - u_2$)越大，重叠面积越小，反映到神经元数据上，也就是信息冗余度越小，即类别间的可分性越大。若均值间的距离相等（图 7-1c 、b），则标准差越大（即测量空间的数据越分散），重叠面积越大，反映到神经元数据上，也就是信息冗余度大，即类别间的可分性越小。

利用公式（7-14）得到尽可能大的距离，则类别的区分度越高。

$$d_{ij} = \frac{|u_i - u_j|}{\sigma_i + \sigma_j} \tag{7-14}$$

接下来根据模型对问题进行求解。

7.3 模型求解

使用搜集到的 5436 组数据，提取 74 个有效的特征属性，编号为 A0，A1，…，A73。接着进行属性选择，借用第一问的结论，选中主要属性 A0，A15，A37，A67。选中的四个属性的物理意义在第 1 问已经详细说明，因此不再赘述。

部分属性选择后的数据如下：

表 7-1 部分数据

编号	A0	A15	A37	A67
1	5	0.36	51.32319	231.0549
2	6	0.36	47.44908	221.7939
3	5	0.36	60.35097	188.8497
4	6	0.36	49.19239	204.2974
5	6	0.36	93.6257	314.8174
6	10	0.25	79.60588	318.2065
7	5	1.45	95.50457	349.6956
8	4	0.41	79.65435	654.4963
9	4	0.12	70.03516	275.1594
10	1	0.36	66.97191	92.01026

这样的数据有 5436 行，用期望最大化方法对其进行聚类[3]。

E 步：取期望。

$$\alpha_{ij} = \frac{f(x_i \mid j,\theta)\varepsilon_j}{\sum_{j=1}^{M} f(x_n \mid j,\theta)\varepsilon_j} \tag{7-15}$$

M 步：求极值。

假定条件似然函数是服从高斯分布，有

$$f(x_i \mid j,\theta) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \frac{-(x_i - \mu_j)^2}{2\sigma_j^2} \tag{7-16}$$

我们现在要求的是极值问题。

$$L(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{M}[\alpha_{ij} \ln f(x_i \mid j,\theta) + \alpha_{ij} \ln \varepsilon_j] \tag{7-17}$$

把公式（7-15），（7-16）代入公式（7-17）中求解，可以求得的估计值。

$$\hat{\theta} = \{\hat{\varepsilon}_j, \hat{\mu}_j, \hat{\sigma}_j \mid j = 1,2,...,M\} \tag{7-18}$$

$\hat{\mu}_j$ 为期望估计，$\hat{\sigma}_j^2$ 为方差估计，$\hat{\varepsilon}_j$ 为平均期望估计。

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N}(\alpha_{ij} x_i)}{\sum_{i=1}^{N}\alpha_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N}[\alpha_{i,j}(x_i - \hat{\mu}_j)^2]}{\sum_{i=1}^{N}\alpha_{i,j}} \tag{7-19}$$

$$\hat{\varepsilon}_j = \frac{\sum_{i=1}^{N}\alpha_{ij}}{N}$$

模型的初始参数 θ^0 可以取随机值。

设置循环终止条件 $T=10^{-6}$，迭代次数 I=100。为保证类别的可区分性，首先计算每一个类别的均值和标准差，再根据公式（求解模型的上面一个 7-14）求解类间距离，最终所有数据最终被聚成了七类。依据我们的模型，聚类结果如下表所示。

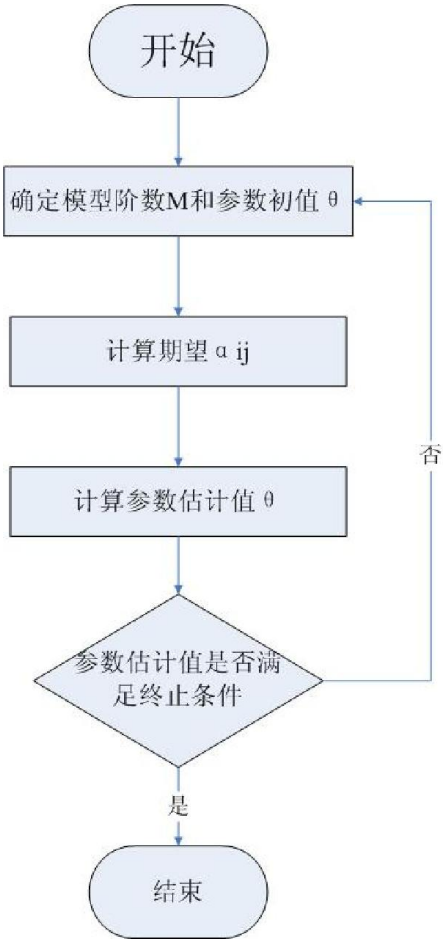


图 7-2 模型流程图

表 7-2 模型运行结果

聚类EM模型		A0		A15		A37		A67	
	先验概率	正态分布均值	标准差	正态分布均值	标准差	正态分布均值	标准差	正态分布均值	标准差
D_0	0.2451	5.9188	2.5756	0.2539	0.1219	78.7835	8.2171	415.2998	235.7664
D_1	0.3572	6.166	1.3319	0.4997	0.1514	66.0803	7.5512	217.1782	50.391
D_2	0.0073	7.3307	1.5161	0.3	0.3798	59.6184	5.6895	9728.6085	3660.4305
D_3	0.0117	1.2355	0.4988	0.9066	0.2544	1.4464	4.0207	255.8724	122.6234
D_4	0.0305	10.5972	10.122	0.2296	0.1523	78.2225	12.0423	1133.9067	579.5806
D_5	0.193	2.9652	1.4455	0.5065	0.3593	88.4572	10.598	184.8649	138.6856
D_6	0.1552	5.1993	2.472	0.844	0.6598	60.8998	11.2129	195.8408	94.2989

下面给出命名方式。

我们首先用附录 AC 中的数据在新分类模型下进行验证，得到七类数据被划分到了其中的四类，还存在新的三个类别。结果如下表所示：

表 7-3 新老类别对照表

D_0	D_1	D_2	D_3	D_4	D_5	D_6
C_2 C_3	未知	未知	未知	C_0	C_1 C_6	C_4 C_5

有表中可知，新类中的 D_0 类对应了老类别中的椎体神经元与二极中间神经元，D_1、D_2、D_3 为新的类别，D_4 对应运动神经元，D_5 对应普肯野神经元与感觉神经元，D_6 对应三极中间神经元与多极中间神经元。

在上述 4 个属性中，搜寻各新类的显著特征，我们进行了以下命名建议：按照 7 中神经元在上述 4 中特征属性的特点，结合神经学知识，尽量将外观特点包含在名称中。同时可以运用我们给出的新老分类方式对照表（表 7-3），结合老类别的特点来总结分析。

我们班门弄斧，尝试了如下命名：

表 7-4 新分类命名尝试

类别	建议名称	理由
D_0	多干放射状神经元	相比别的指标，其"比较扩大角度"平均值显著较大，也就是线条较直。且干数较多。
D_1	普通神经元	35%的数据在这个分类中，其没有显著特征，为大多数的情况。命名方式可以任意。
D_2	多干长须神经元	其"段路径长度"最大值格外突出，比一般的类别长很多。
D_3	毛线状神经元	其"比较扩大角度"平均值显著较小，说明歪曲程度较高。
D_4	多干中须神经元	干的数目显著较多，而且其"段路径长度"最大值也较大。
D_5	珊瑚状神经元	比较普通，结合题目中的 5 中分类。
D_6	多极中间神经元	比较普通，结合题目中的 5 中分类。

8 问题 4——比较分析不同物种的同类神经元形态特征

8.1 问题分析

本问题要求使用前面建立的模型，因此我们使用朴素贝叶斯模型。要确定不同动物神经系统中同一类神经元的形态特征，只要改变一下训练实例集，让朴素贝叶斯重新训练一下即可[4]。本题的训练实例集可以选择不同动物的相同神经元作为训练实例集。附录 A 中提供了 3 个猪的普肯野神经元和 3 个鼠的普肯野神经元，同时，我们又搜集了 Camero 实验室的 20 组猫的脊髓运动神经元数据和 17 组鼠的脊髓运动神经元数据，以及 Lweis 实验室的 20 组人类椎体神经元数据和 20 组猴子的椎体神经元数据。我们希望可以借此比较分析出不同物种的同类神经元形态特征。

8.2 问题求解

8.2.1 对比猪和鼠的普肯野神经元：

我们首先画出猪和鼠的神经元属性分布图，对比结果如下图所示。

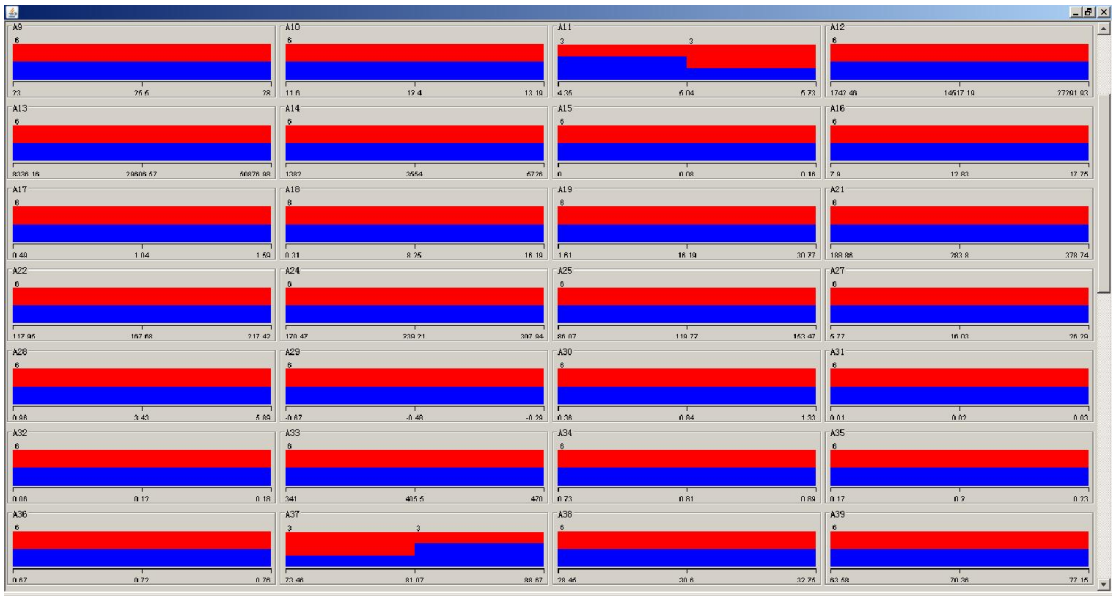


图 8-1 猪和鼠的普肯野神经元属性分布情况
（蓝色代表猪，红色代表鼠）

从图 8-1 可以看出，除 A_{11}（个体整体标准差）和 A_{37}（比较扩大角度平均值）两个属性有明显区别外，其他属性特征十分接近。因此，从这 74 个属性上不能绝对区分猪和鼠的普肯野神经元的异同，不能建立一个准确的分类器模型。

但是我们可以从分布图上看出，对于 A_{11} 属性，猪的大部分值在 4.35-5.04 之间，而鼠的多在 5.04-5.73 之间；对于 A_{37} 属性，猪的大部分值在 81.07-88.67 之间，而鼠的多在 73.46-81.07 之间。也就是说，在个体整体标准差（standard

deviation）上，猪的平均比鼠的小；而在“比较、扩大、角度”平均值（Avg comp. ampl. angle）上，猪的平均比鼠的大。

但是，基于统计观察模型只能给出关系数据，却不能定量分析。因此，我们从网上搜集了 77 组相关数据，并建立模型。

8.2.2 对比猫和鼠的脊髓运动神经元

Step 1：属性选择。首先用问题 1 的方法对猫和鼠的脊髓运动神经元数据（猫 20 组，鼠 17 组）进行属性选择，得出 A_{15} 可以作为猫和鼠的脊髓运动神经元唯一指标，也就是最小直径（Minimum diameter）。

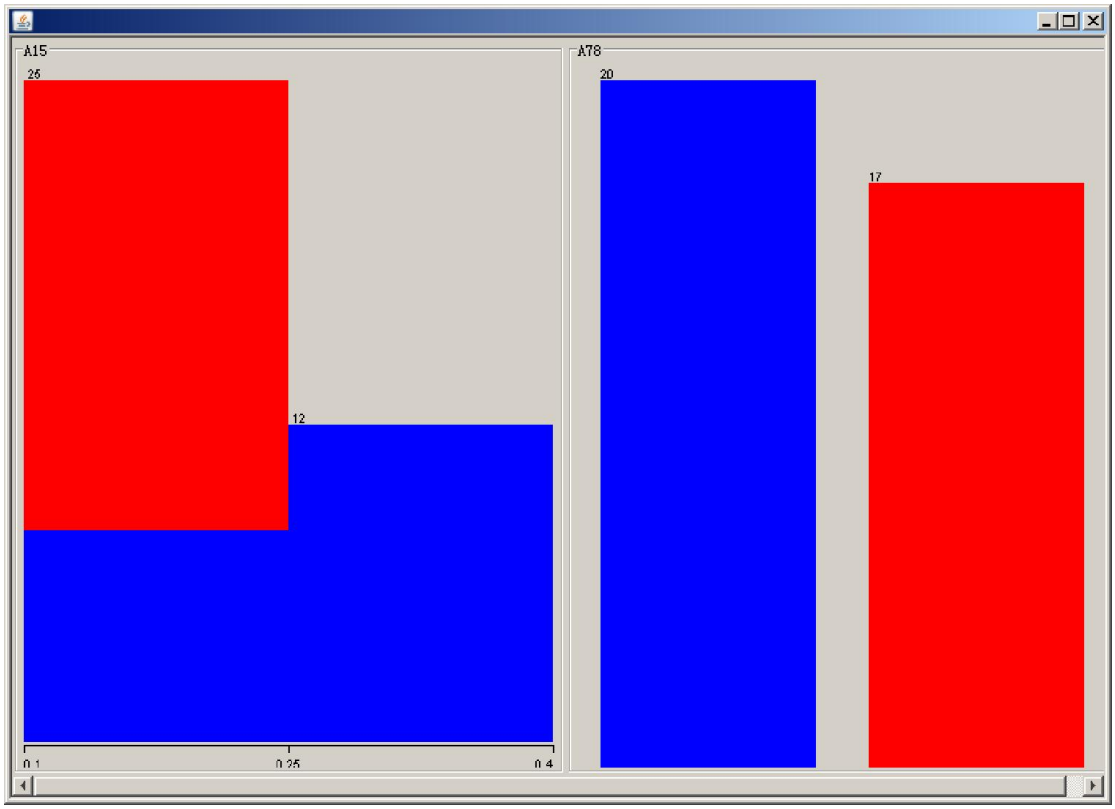


图 8-2 猫和鼠的脊髓运动神经元 A_2 属性的分布图
（蓝色为猫，红色为鼠）

从图中我们可以看出很明显鼠的脊髓运动神经元的最小直径平均小于猫的。

Step 2：朴素贝叶斯分类器的建立。用 A_{15} 建立朴素贝叶斯分类器，模型如下表所示：

表 8-1 猫和鼠的脊髓运动神经元朴素贝叶斯模型

猫的脊髓运动神经元	概率	0.54
	平均值	0.2679
	标准差	0.1204
鼠的脊髓运动神经元	概率	0.46
	平均值	0.1286
	标准差	0.0071

Step 3：十折交叉验证。详情见附录 3。

表 8-2 部分验证的正确率

编号	猫概率	鼠概率	实际类别	预测类别	正确率
1	*1	0	猫	猫	100%
2	*1	0	猫	猫	
3	0.08	*0.92	鼠	鼠	
4	0.08	*0.92	鼠	鼠	

通过贝叶斯模型和 A_{15}（最小直径）属性，猫和鼠脊髓运动神经元的分类正确率可以达到 100%。可见猫鼠的脊髓运动神经元最小值经为主要区分特征，且此特征在朴素贝叶斯模型下分类显著。

8.2.3 对比猴子和人类的椎体神经元

Step 1：属性选择。首先用问题 1 的方法对猴子和人类的椎体神经元数据（各 20 组）进行属性选择，得出 A_{50} 和 A_{76} 可以作为猫和鼠的脊髓运动神经元特征指标，也就是 P/C 直径标准差（P/C diameter standard deviation）和平均路径距离（Avg path distance）。

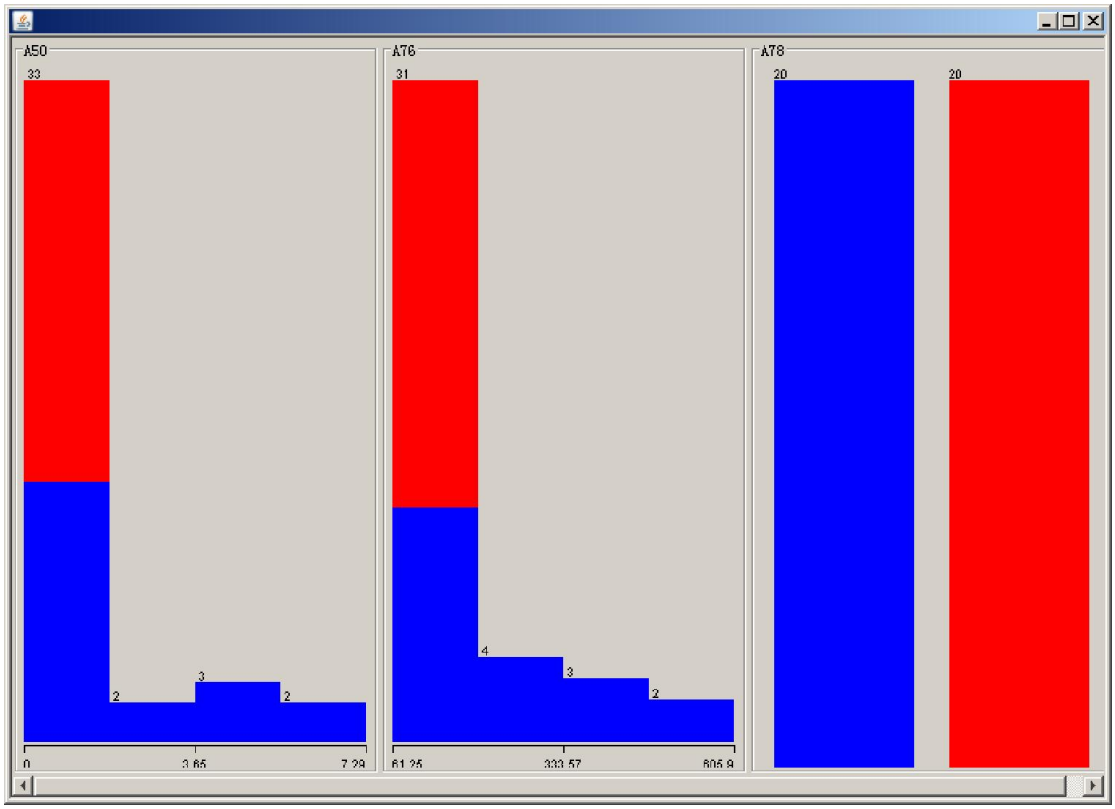


图 8-3 猴子和人类的椎体神经元 A_{50}，A_{76} 属性的分布图
（蓝色为猴子，红色为人类）

从图中我们可以很明显的看出人类的椎体神经元的 P/C 直径标准差和平均路径距离平均小于猴子的。

Step 2：朴素贝叶斯分类器的建立。用 A_{50}，A_{76} 建立朴素贝叶斯分类器，模型如下表所示：

表 8-3 猴子和人类的椎体神经元朴素贝叶斯模型

猫的脊髓运动神经元	概率		0.5
	A_{50}	平均值	1.8228
		标准差	2.3355
	A_{76}	平均值	224.15
		标准差	153.71
鼠的脊髓运动神经元	概率		0.5
	A_{50}	平均值	0.4375
		标准差	0.1648
	A_{76}	平均值	94.267
		标准差	13.157

Step 3：十折交叉验证。详情见附录 4。

表 8-4 部分验证的正确率

编号	猫概率	鼠概率	实际类别	预测类别	正确率
1	0.008	*0.992	猴	猴	100%
2	0.013	*0.987	猴	猴	
3	*1	0	人	人	
4	*1	0	人	人	

通过 P/C 直径标准差（P/C diameter standard deviation）和平均路径距离（Avg path distance）属性，猴子和人类的椎体神经元的分类正确率可以达到 97.5%。可见猴子和人类的椎体神经元 P/C 直径标准差和平均路径距离为主要区分特征，且此特征在朴素贝叶斯模型下分类比较显著。

8.2.4 对比 6 个种类的不同物种的不同神经元

为了研究物种对神经元几何特征分类的影响，我们选取猫和鼠的脊髓运动神经元、猴子和人类的椎体神经元以及猪和鼠的普肯野神经元等，从物种和外形特征上共分为 6 种，分别定义为 C_1'，C_2'，C_3'，C_4'，C_5'，C_6'。将这 6 种进行特征分析并建立分类模型。

Step 1：属性选择。首先用问题 1 的方法对 6 组神经元数据进行属性选择，得出 A_{16} 和 A_{28} 可以作为 6 组神经元特征指标，也就是最大直径(Maximum diameter)和平均比较距离（Avg comp. length)。

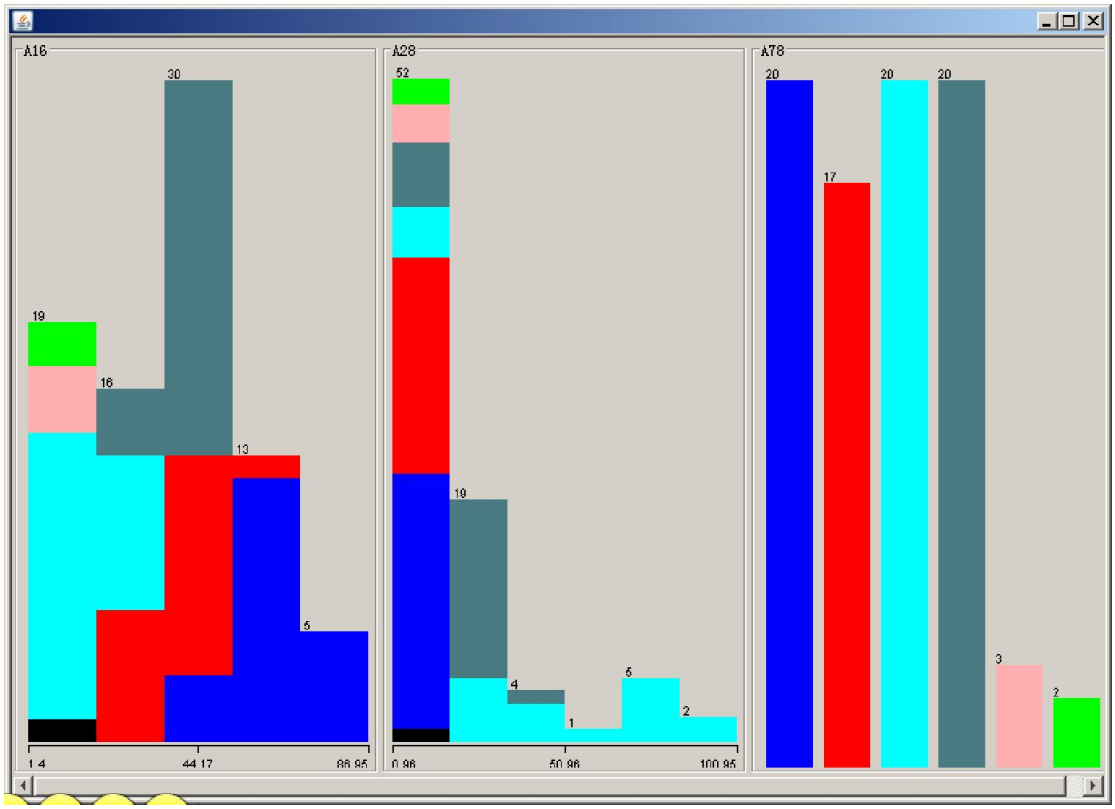


图 8-4 6 组神经元数据 A_{16}，A_{28} 属性的分布图

（蓝色为猫脊髓运动神经元，红色为鼠脊髓运动神经元，青色为猴椎体神经元，灰色为人椎体神经元，粉丝为猪的普肯野神经元，绿色为鼠的普肯野神经元）

Step 2：朴素贝叶斯分类器的建立。用 A_{16}，A_{28} 建立朴素贝叶斯分类器，模型如下表所示：

表 8-3 6 组神经元朴素贝叶斯模型

C_1'	概率		0.24	C_4'	概率		0.24
	A_{16}	平均值	63.791		A_{16}	平均值	39.3943
		标准差	13.168			标准差	3.117
	A_{28}	平均值	9.3286		A_{28}	平均值	21.8277
		标准差	2.0139			标准差	5.2859
C_2'	概率		0.2	C_5'	概率		0.05
	A_{16}	平均值	38.705		A_{16}	平均值	16.8984
		标准差	10.629			标准差	0.8623
	A_{28}	平均值	3.5148		A_{28}	平均值	1.6259
		标准差	0.7101			标准差	0.5748
C_3'	概率		0.24	C_6'	概率		0.03
	A_{16}	平均值	16.793		A_{16}	平均值	8.4492
		标准差	4.1298			标准差	0.176
	A_{28}	平均值	47.009		A_{28}	平均值	5.4874
		标准差	29.503			标准差	0.6097

Step 3：十折交叉验证。详情见附录 5。

表 8-4 部分验证的正确率

编号	C_1'	C_2'	C_3'	C_4'	C_5'	C_6'	实际类别	预测类别	正确率
1	*1	0	0	0	0	0	C_1'	C_1'	100%
2	*1	0	0	0	0	0	C_1'	C_1'	
3	0	0	0	*1	0	0	C_4'	C_4'	
4	0	0	0	*1	0	0	C_4'	C_4'	
5	0.008	0	*0.992	0	0	0	C_3'	C_3'	
6	0	0	*1	0	0	0	C_3'	C_3'	
7	0	*0.995	0.005	0	0	0	C_2'	C_2'	
8	0.074	*0.926	0	0	0	0	C_2'	C_2'	
9	0	0	0	0	0	*1	C_6'	C_6'	

通过最大直径（Maximum diameter）和平均比较距离（Avg comp. length）属性，6 组神经元的分类正确率可以达到 97.561%。可见 6 组中最大直径和平均比较距离为主要区分特征，且此特征在朴素贝叶斯模型下分类比较显著。

综上所述，朴素贝叶斯分类器，对不同物种不同神经元有比较好的区分度。

9 问题 5——预测神经元生长变化

9.1 问题分析

附录给的样本神经元的数据是通过实验的手段，对神经元进行染色，然后观察得到。但是神经元在被染色之后，很有可能会衰亡，技术上可以认为无法染色得到同一个神经元的各个生长状态的数据[5]。

所以，预测神经元形态的生长变化，需要对大量同一类型的神经元数据进行分析。我们决定采用聚类的方法，对大量同一类型的神经元的几何形态特征进行 EM 算法聚类，由于相同神经元的不同特征反应在该神经元的几何形态变化上，可以认为处于不同生长时期的神经元有不同形态。(**假设：**从 neuromorho.org 网站上得到的 1908 个人类椎体神经元数据是分布在该类别神经元生长发育的各个阶段的。)

聚类后产生 n 个类，可以认为该物种神经元分为 n 个生长阶段，有了这个聚类模型后，可以根据神经学的知识进行特征分析，判断出 n 个类在神经元生长阶段的先后顺序，从而能够判断某一神经元处于哪个生长阶段，并对其下阶段的生长状况进行预测。

9.2 模型的建立与求解

9.2.1 聚类分析

首先对 1908 个人类椎体神经元在所有属性（未经过特征筛选的 78 个）上进行 EM 聚类。模型的建立过程详见问题 3。

聚类后我们得到 6 个类别的数据，暂且定义为 P1, P2, P3, P4, P5, P6。目前，我们还不知道这六个类别的具体含义，只知道它们是 6 个不同的年龄段，但先后顺序不知。比例也比较均匀，详情参见下图。

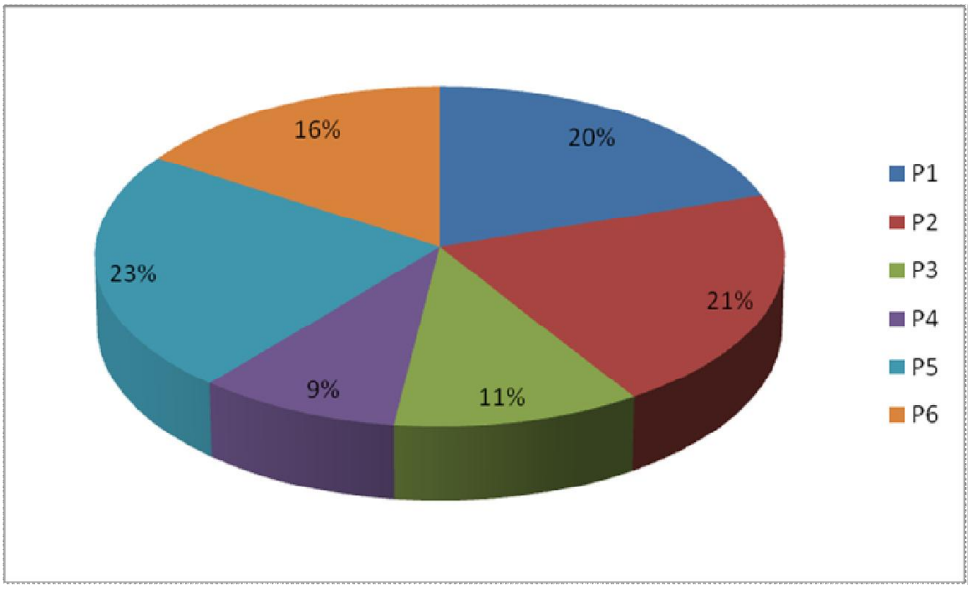


图 9-1 各成长阶段分布饼图

对于 P1-P6 聚类后，将其类别标记在个体的数据中，从而可以进行下一步的属性选择和分析实验。

9.2.2 特征提取

拥有了类别标记，可是具体类别代表的成长阶段仍然是未知的，但在 77 个属性中，很难统计出成长过程。我们首先需要对这 77 个属性进行特征提取，然后再针对提取的少量属性进行统计总结。

特征提取模型详见问题 1。经过问题 1 中方法的特征提取后，我们得到。A_3，A_4，A_5，A_6，A_7，A_{12}，A_{13}，A_{14}，A_{21}，A_{22}，A_{24}，A_{27}，A_{28}，A_{33}，A_{37}，A_{39}，A_{41}，A_{42}，A_{45}，A_{50}，A_{51}，A_{56}，A_{57}，A_{65}，A_{67}，A_{71}，A_{72} 等 27 个特征属性。（各属性代表的具体含义详见附录 1）

9.2.3 贝叶斯分类模型建立

通过上述特征属性和类别，我对这些数值建立贝叶斯分类模型，建模方法详见问题 1。

但是在贝叶斯模型建立之前，我们需要注意，本题的模型建立方法虽然与问题 1 相同，但是其目的不同。问题 1 是问了分类，而本问是为了每个属性的统计特征而进行比较。所以，本问在建立模型前，首先要进行数据的标准化。

用 SPSS16.0 对这 27 个属性进行标准化，用标准化后的数据建模后得到 6 个类别的这 27 个属性的均值和标准差，详情见附录 6。下表为 P1 类别的 27 个类别统计特征举例。

表 9-1 P1 类别的 27 个类别统计特征举例

属性编号	均值	标准差	属性编号	均值	标准差
ZA3	-0.5055	0.7994	ZA37	0.035	1.0909
ZA4	0.4086	0.717	ZA39	0.1132	1.1232
ZA5	-0.4631	0.7303	ZA41	-0.3635	1.1121
ZA6	0.2708	0.6975	ZA42	0.1522	1.2748
ZA7	-0.2236	0.6272	ZA45	0.0147	2.1947
ZA12	-0.7183	0.4822	ZA50	-0.0868	2.1326
ZA13	-0.6649	0.5955	ZA51	-0.5914	0.9674
ZA14	-0.5594	0.5784	ZA56	0.1119	2.2045
ZA21	-0.7229	0.5065	ZA57	-0.3467	0.7622
ZA22	-0.7376	0.5619	ZA65	-0.5228	0.7983
ZA24	-0.7297	0.5316	ZA67	-0.5717	0.4941
ZA27	-0.314	0.7109	ZA71	-0.5932	0.6025
ZA28	-0.0778	1.6251	ZA72	-0.6507	0.8467
ZA33	-0.5333	1.1899			

得到的特征值分别是 P1-P6 的 27 个属性标准化后的平均值和标准差。

平均值可以看出各类的 27 个属性的数值变化情况，标准差可以看出 27 个属性的离散程度。离散程度越高，表示包含的信息量越大。

9.2.4 成长期排序

我们已经聚成了 6 个类别，接下来应该对其进行排序就，根据序列大小判断每一个类别的年龄。考虑到数据的离散程度越高，信息量越大的特性，按照如下公式计算每一个类别的权重值：

$$value_j = \sum_{i=1}^{27} \alpha_i \mu_i \sigma_i \tag{9-1}$$

其中 $value_j$ (j=1, 2, …, 6)是类别的权重。a_i 取-1 和 1 两个值，如果 μ_i 负增长，我们就取-1。μ_i 代表第 i 个属性的平均值，而 σ_i 代表第 i 个属性的标准差。

根据参考文献和常识，我们知道神经元的生长是树突和轴突的增长，故而其体积、面积等数据应该变大，而最大 x y z 值等都应变大，最小 x y z 值等都应变小。

根据分析，以及从网络上询问专家、查阅文献，我们构造了如下 a_i 值表：

表 9-2 a_i 取值列表

属性	a_i 值	属性	a_i 值	属性	a_i 值
A3	1	A22	1	A45	1
A4	-1	A24	1	A50	1
A5	1	A27	1	A51	1
A6	-1	A28	1	A56	1
A7	1	A33	1	A57	1
A12	1	A37	1	A65	1
A13	1	A39	1	A67	1
A14	1	A41	1	A71	1
A21	1	A42	1	A72	1

通过计算，得到 P1-P6 的成长排列值：

表 9-3 成长度排序值列表

P1	P2	P3	P4	P5	P6
-6.97861687	2.89814065	-1.54801011	31.80469804	-0.01727021	-0.35405473

我们拟合出人类椎体神经元的成长曲线，如下图所示。其成长阶段依次为 P1 ,P3,P6,P5,P2,P4。

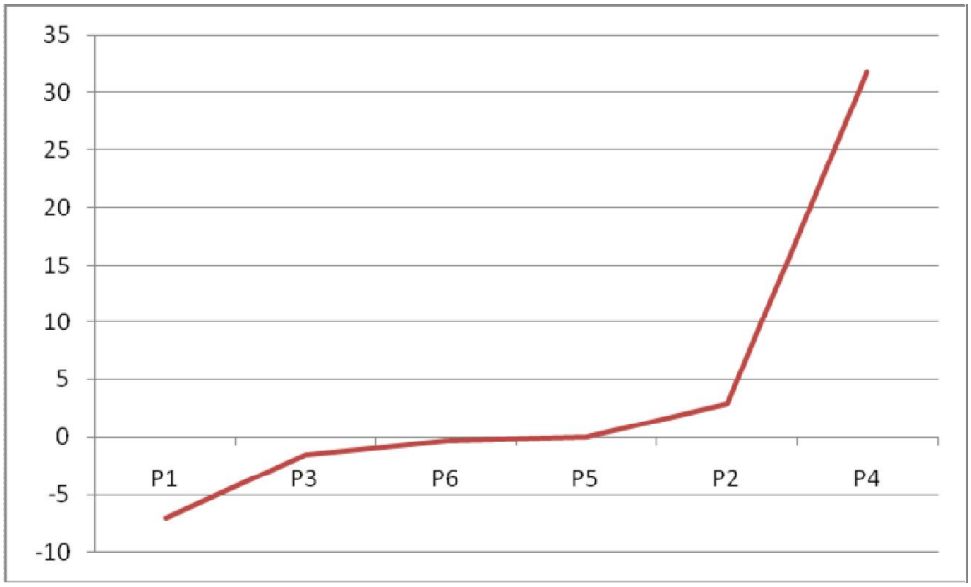


图 9-2 人类椎体神经元的成长曲线

也就是说，我们把预测问题转化成了分类问题，如果知道了某神经元形态特征，我们便可以判别它的生长阶段，且我们知道后面的阶段是什么，故达到预测的目的。

例如一个神经元被判别到 P3 阶段，我们变可以预测它将会向 P6 阶段生长。

9.2.5 预测模型验证

我们对这 1908 个数据进行十折交叉验证（详见问题 1），分类的正确率达到 82.5996 %，由于版面问题，这里不再举例说明，具体数据见附录 7。

表 9-4 10 折交叉验证结果

次数	正确率	次数	正确率
第一次	81.15%	第六次	81.15%
第二次	81.15%	第七次	82.20%
第三次	81.68%	第八次	80.10%
第四次	90.05%	第九次	84.74%
第五次	82.20%	第十次	86.32%

10 模型的评价与改进

10.1 创新点与优势

在本题的解决过程中，我们遇到了不少的困难，但最后都被我们一一解决了，在论文的最后，总结我们的创新点和优势如下：

1.比较好的解决了题目中的问题。分类的精度均在 96%以上，而聚类的精度也基本保持在 80%以上。特征的提取也做到了少而精，结果容易理解。

2.用计算机的编程方法处理了海量的数据（整个 neuromorpho.org 网站数据库的所有数据）。保证了数据的完整性，使得结果比较客观。

3.问题 1 到 4 中推导并优化了经典的数学模型。针对朴素贝叶斯分类模型、最好优先选择模型、期望最大化聚类模型以及遗传规划模型，在本题的应用中都进行了不同程度的改进和创新。

4.在问题 5 中提出了分阶段判定预测神经元的生长。避免了一般的回归方法上数据与背景知识的不足，建立了新的预测模型。

5.用专业的神经元研究软件 Neuron 对原始数据进行了属性提取。通过编程，我们批量提取了大量原始数据的几何特征属性，用 Neuron 可以统计出的神经元几何特征属性高达 78 个之多。

10.2 不足与改进

但是由于时间仓促，在建立的模型上还存在一些问题。

1.对于问题 2 的模型。问题 2 的模型已经比较好的完成了问题 2 的任务，但是如果有大量新类别数据涌入的时候，还不能将新类别立刻分门别类。

可以采用分类和聚类相结合的模型，已知的分类器将新类别拒绝后，再将其进行聚类，这样未知类别的数据也可以做到分门别类。再对其分别进行特征提取后，可以更方便更直观的命名。

2.对于问题 5 的模型。问题 5 的模型虽然可以分阶段预测，但还不能够拟合出连续的生长曲线。

可以先用本题针对问题 5 的预测神经元的成长阶段，在每个阶段中，再提取对应阶段的特征属性进行回归拟合，这样可以比较科学的进行曲线预测。例如，神经元成长的第一阶段树突生长迅速，但轴突生长缓慢，而第二阶段反之的话，用单一的评价标准就不科学了。

11 参考文献

[1] 俞立平等，学术期刊评价中主成分分析法应用悖论研究，情报理论与实践，32(9)：84-87,2009。
[2] 朱周华，期望最大(EM)算法及其在混合高斯模型中的应用，现代电子技术，24：88-90，2003。
[3] 王平波等,混合高斯概率密度模型参数的期望最大化估计,声学技术,26(3)：498-502，2007。
[4] 张敏等,基于朴素贝叶斯分类器的大鼠体态自动识别,航天医学与医学工程，18(4):370-374，2005。
[5] Irwin B.Levitan 等，细胞和分子生物学，北京：科学出版社，2008。